generic HMM for multi-parent populations

Karl Broman
Biostatistics & Medical Informatics, UW–Madison

@kwbroman
kbroman.org
github.com/kbroman
kbroman.org/Talk_GenericHMM

These are slides for a talk for the CTC (www.complextrait.org/ctc2021/) on 1 Sept 2021.

Slides: kbroman.org/Talk_GenericHMM/generic_hmm.pdf

Slides with notes: kbroman.org/Talk_GenericHMM/generic_hmm_notes.pdf

Source: github.com/kbroman/Talk_GenericHMM

Related paper on bioRxiv: doi.org/gswx
Multi-parent populations are experimental crosses starting from multiple inbred founder lines.

Major examples include the Collaborative Cross, a set of 8-way recombinant inbred lines in mouse, and Heterogeneous Stock, which have been developed in both mice and rats and are advanced intercross populations derived from 8 founders. The Diversity Outbred mouse population is similar to HS. In plants, multi-parent recombinant inbred lines are called MAGIC lines (for multiparent advanced generation inter-cross).

The offspring chromosomes will be mosaics of the founder chromosomes. Multi-parent populations can be homozygous (like RIL) or heterozygous (like HS). The number of founders need not be 8.
A key step in the analysis of multi-parent populations is genome reconstruction: using dense SNP genotypes in the founders and MPP offspring to infer the haplotypes across the genome.

Here we consider a 1 Mbp region on chromosome 14 in a single Diversity Outbred Mouse. Open and closed circles indicate AA and BB genotypes at SNPs. Gray circles indicate AB heterozygous genotypes. Using the SNP data along the chromosome, we can calculate the probability of each possible genotype at each position.

For this mouse, the left half of the interval looks to be homozygous DD, while the right half looks to be heterozygous AD.
One could skip the whole genome reconstruction and just do QTL analysis at the SNPs, as is done in GWAS. If the genotyped SNPs include individual causal polymorphisms, this could be best.

But if there are multiple causal polymorphisms in a region QTL analysis with the inferred haplotypes may be more powerful. Moreover, if the founder strains have been sequenced, you can use the reconstructed genomes to get inferred genotypes at all polymorphisms in the founders. (Similar approaches were used in human GWAS, based on HapMap SNPs.)

Here, the single-SNP analysis shows significant evidence for a single QTL on chromosome 1. The haplotype analysis indicates evidence for a second QTL on chromosome 4.

Beyond QTL mapping, genome reconstructions are useful in data diagnostics. For example, the estimated number of crossovers is useful when assessing sample quality.
Here is the reconstructed genome of a Diversity Outbred mouse. (The white segments are undetermined.)

Our goal is to figure this out, using SNP genotypes on this mouse plus the 8 founder lines.
The main approach for genome reconstruction is to use a hidden Markov model. The underlying diplotypes we’re trying to determine follow a Markov chain \( \{G_i\} \), but are unobserved. We observe SNP genotypes \( \{O_i\} \), with an assumed conditional independence structure, where given \( G_i \), \( O_i \) is conditionally independent of everything else.

Three sets of parameters govern the model: the initial and transition probabilities, which concern the pattern of underlying genotypes on the MPP chromosomes; and the emission probabilities, which relate the underlying genotypes to the observed SNP genotypes and largely concern a model for SNP genotyping errors.
I’ve spent quite a lot of time studying the pattern of genotypes on MPP chromosomes, first with a paper on multi-way recombinant inbred lines, but then following up with three further papers considering extra generations of outbreeding, the genotypes at intermediate generations, and the patterns in advanced intercross populations such as Diversity Outbred mice.

The mathematics is interesting but tedious. And is it necessary? It would be nice to have a generic approach that could be used generally.
And that is what I propose here. Imagine a population of $k$ founders in known (but not necessarily equal) proportions, and that a multi-parent population is formed by random mating for $n$ discrete generations. In this case, we can calculate the transition probabilities exactly.

We could apply these equations more generally. We need just specify the proportions of the founders (which should be known from the design of the study) and the effective number of generations of random mating. The latter might be calibrated by considering the map expansion (the proportional increase in the number of recombination breakpoints, relative to a single meiosis). This could be approximated by computer simulation.

For a heterozygous population, like HS or the DO, we draw two random chromosomes. For a homozygous population, like MAGIC lines or the Collaborative Cross, we can pretend that they are doubled haploids, with a single random chromosome like above.

For the X chromosome, we can use the same equations, replacing $n$ with $\left(\frac{2}{3}\right)n$, due to recombination only happening on the X chromosome in females.
If we apply our approach to data from Diversity Outbred mice, the results with the generic model proposed above are basically identical to the use of the more-exact model. For data from Al-Bargouthi et al (2021), this is the biggest difference seen: the LOD curves are not distinguishable, as the biggest difference is just 0.01.
Reconstructions of the genomes of Collaborative Cross lines are identical for autosomes, but there are important differences for the X chromosome.

This slide shows the reconstruction of the X chromosome in Collaborative Cross line CC038, but the exact model (top panel) and by the approximate model (bottom panel).

The analysis differs in that the top model excludes three of the eight founders and weighs one of the other five more highly.

These results differ in a region around 135 Mbp, where in the bottom panel, B6 and NOD are assigned equally probability, as they are identical in the region, but the top panel was able to exclude B6.
The X chromosome in the Collaborative Cross behaves different than autosomes. We list the crosses female × male; note that the Y chromosome comes from the H strain and the X chromosome comes from the five strains A, B, C, E, and F, with the average proportion from the C strain between twice that of the others.

This can be really useful information (provided that it is correct), particularly as the X chromosome shows reduced polymorphism compared to the autosomes. Many of the CC founders share large stretches of DNA on the X chromosome.
Summary

- Generic model for genome reconstruction in multi-parent populations
- Specify relative proportions of founders
  + effective number of generations of random mating
- Basic conclusion: HAPPY is effective
- Implemented in R/qtl2 as cross types genril\textsubscript{n} and genail\textsubscript{n}
  (replacing \textit{n} with the number of founders)
- bioRxiv manuscript: doi.org/gswx

It's always good to provide a summary.
Here’s where you can find me and these slides, as well as a preprint giving further details on the work.